Banff Challenge 2: Stanford Entry

B. Efron, T. Hastie, O. Muralidharan, B. Narasimhan, J. Scargle,
Robert Tibshirani, Ryan Tibshirani

December 10, 2010

Problem 1

We observe n "marks" $x_1, x_2, \ldots, x_n$ in the unit interval, $N_{back} = A/10$ from the background exponential density

$$f_0(x) = 10e^{-10x}$$

and $N_{sig} = D\sqrt{2\pi\sigma^2} = 0.0752D$ from a normal density with unknown mean E and known variance $\sigma = 0.03$,

$$f_1(x; E) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-E}{\sigma}\right)^2}.$$

We want to test the null hypothesis $H_0 : D = 0$ versus the alternative $H_1 : D > 0$, and also to provide estimates of D and E and standard errors of these estimates.

The situation above can be thought of as observing an independent sample of size n from the mixture density

$$f(x; D, E) = \left(\frac{n - cD}{n}\right) f_0(x) + \frac{cD}{n} f_1(x; E) \tag{1}$$

where $c = \sqrt{2\pi\sigma^2} = 0.0752$. Let $\left(\hat{D}, \hat{E}\right)$ the maximum likelihood estimates of (D, E) in the model. Then we can test the null hypothesis H_0 using the maximum likelihood ratio statistic

$$L = \frac{\prod_{i=1}^n f\left(x_i; \hat{D}, \hat{E}\right)}{\prod_{i=1}^n f_0(x_i)}, \tag{2}$$

rejecting H_0 for large values of L. In our submission, "large" was defined by simulation of L under the null hypothesis, rather than from theoretical chi-squared calculations.

"Lindsey's method" [1] allows a restatement of the problem in terms of Poisson regression: we bin the data, say into 100 bins of width $d = 0.01$ on the unit interval, letting

$$y_k = \#\{x_k \textit{ values in } k^{th} \textit{ bin}\}.$$

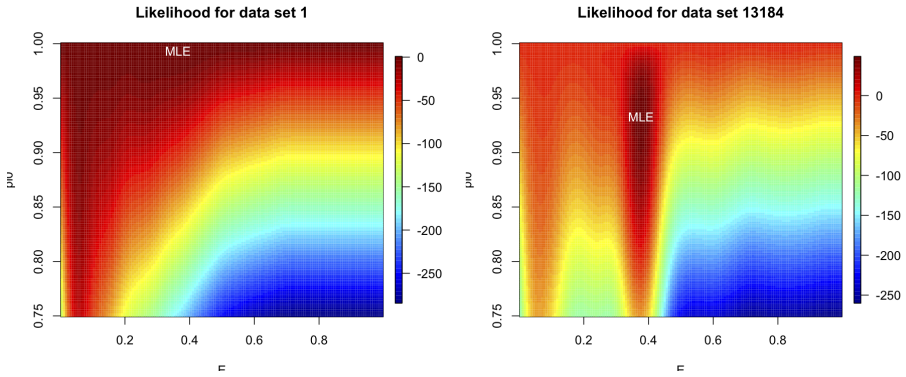

Figure 1: Likelihood surfaces for a typical null (left) and non-null (right) data sets. E varies along the x-axis, and the null proportion $\pi_0 = \frac{n-cD}{n}$ varies on the y-axis.

Define $\mu_k(D, E)$ to be the expected value of y_k under model 1, approximately

$$\mu_k(D, E) = nd \cdot f(m_k; D, E),$$

where m_k is the midpoint of bin k. Then the independent Poisson model for the counts y_k,

$$y_k \sim Poisson(\mu_k(D, E)), \tag{3}$$

yields almost the same maximum likelihood estimates $\left(\hat{D}, \hat{E}\right)$ as before, the difference quickly becoming negligible as the bin width d goes to zero.

Model 3 is a two-parameter nonlinear Poisson regression. Figure 1 shows two typical likelihood surfaces. The likelihood function (equation 1) can be multimodal, so optimizing it requires some care. We found the maximum likelihood estimates $\left(\hat{D}, \hat{E}\right)$ using a simple grid search, and used the nonparametric bootstrap to estimate the variability of our estimates.

Figure 2 shows the log likelihood ratios for the 20,000 supplied data sets, plotted against the estimated signal location $\hat{E}$. About 11% of the data sets show signal at the 1% significance level. The figure suggests that the non-null data sets were generated using 6 or 7 distinct values of E. In the three detection scenarios $(D, E) = (1010.0, 0.1)$, $(137.0, 0.5)$, $(18.0, 0.9)$, our test has power 34.83%, 43.35%, 1.75% at the 1% significance level. The third scenario is particularly challenging. Taking $D = 18.0$ corresponds to observing a single non-null point. Even though the non-null point has a high mean, $E = 0.9$, it is still difficult to distinguish this situation from the null: given 1000 marks from the null f_0, we have a 12% chance of seeing at least one mark above 0.9.

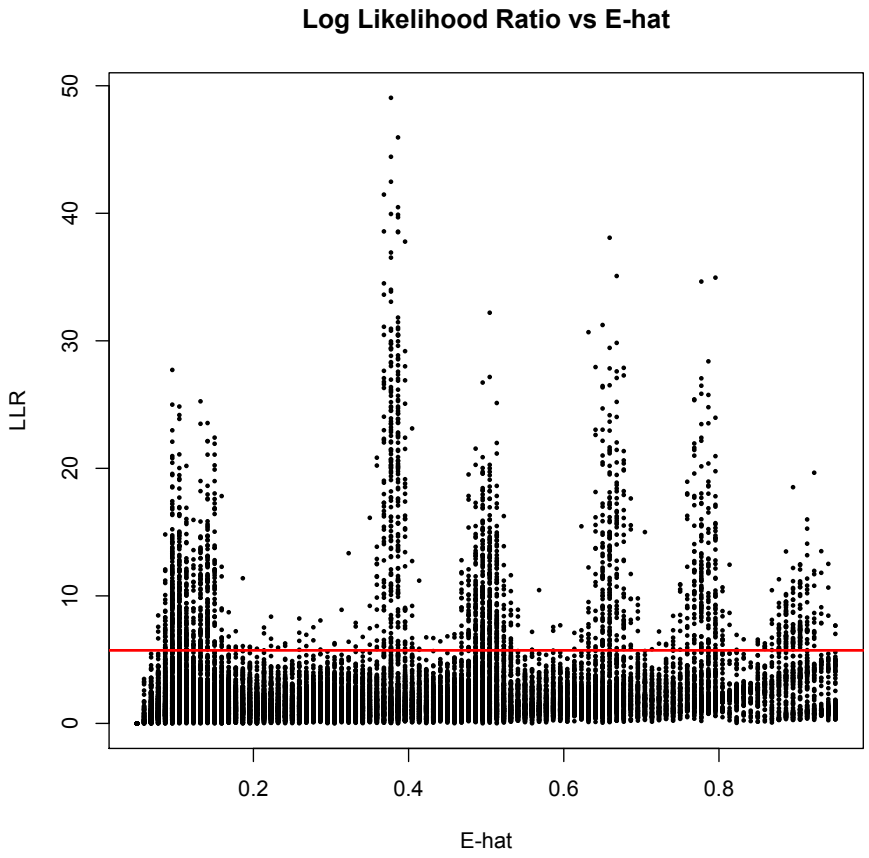

Figure 2: Log likelihood ratios for the 20,000 data sets, plotted against the estimate signal location $\hat{E}$. The red line shows the 1% significance rejection threshold, obtained through independent simulations.

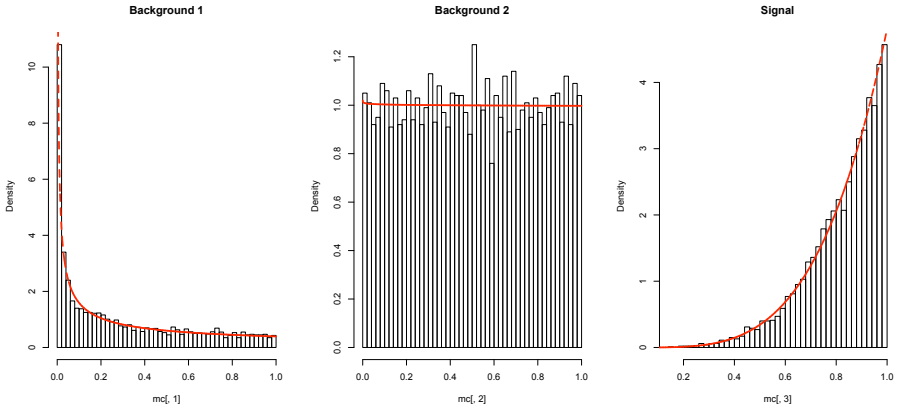

Figure 3: Histogram of background, signal draws, with approximating Beta distributions.

Problem 2

We again observe n marks $x_1, \ldots, x_n$ in the unit interval, n_1 from "background 1," n_2 from "background 2," and n_3 from the signal. We are not given the parametric forms of the background and signal distributions, but are given 5000 draws from each. We wish to test the null hypothesis $H_0 : n_3 = 0$ against the alternative $H_0 : n_3 > 0$.

Although the background and signal distributions are unknown, we can estimate them using the draws that we are given. We estimated the background and signal densities using Beta distributions. Figure 3 shows the draws from the three distributions and the fitted Beta distributions.

We can again think of this problem as observing n independent samples from the mixture density

$$f\left(x; n_1, n_2, n_3\right) = \frac{n_1}{n} f_1\left(x\right) + \frac{n_2}{n} f_2\left(x\right) + \frac{n_3}{n} f_3\left(x\right), \tag{4}$$

where f_1 and f_2 are the background 1 and 2 densities, and f_3 is the signal density. Suppose we have maximum likelihood estimates of n_i, $\left(\hat{n}_1^0, \hat{n}_2^0\right)$, under the null hypothesis $n_3 = 0$, and estimates $\left(\hat{n}_1, \hat{n}_2, \hat{n}_3\right)$ without any restriction on n_3. We can test the null hypothesis H_0 using the generalized likelihood ratio statistic

$$L = \frac{\prod_{i=1}^n f\left(x_i; \hat{n}_1, \hat{n}_2, \hat{n}_3\right)}{\prod_{i=1}^n f\left(x_i; \hat{n}_1^0, \hat{n}_2^0, 0\right)},$$

rejecting for large values of L. In our submission, we defined "large" by simulating from the null, using the distributions for n_1 and n_2 given in the problem statement and the estimated Beta distributions.

The mixture density 4 is easier to fit than the mixture model for problem 1, since it has no unknown nuisance parameters. We found the null MLEs $\left(\hat{n}_1^0, \hat{n}_2^0\right)$

and the unrestricted MLEs $(\hat{n}_1, \hat{n}_2, \hat{n}_3)$ using the standard EM algorithm for mixture models.

About 13% of the 20,000 supplied data sets show signal at the 1% significance level. Our test has 84% power at the 1% significance level in the problem's power testing scenario, with $n_1 \sim \mathcal{N}\left(900, 100^2\right)$, $n_2 \sim \mathcal{N}\left(100, 100^2\right)$, truncated at zero, and "an expected total signal rate of 75 events", which we took to mean $n_3 \sim Poisson\,(75)$.

References

[1] Bradley Efron and Robert Tibshirani. Using specially designed exponential families for density estimation. *Annals of Statistics*, 24(6):2431–2461, 1996.